

SVM and Kernel machine

Stéphane Canu
stephane.canu@litislab.eu

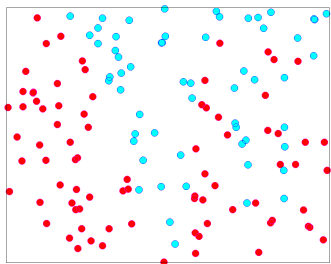
Escuela de Ciencias Informáticas 2011

July 27, 2011

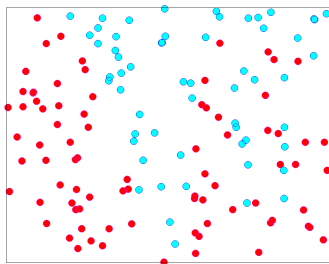
Plan

- 1 Statistical learning and kernels
 - The statistical learning problem
 - Kernels
 - Kernel and hypothesis set
 - Functional differentiation in RKHS

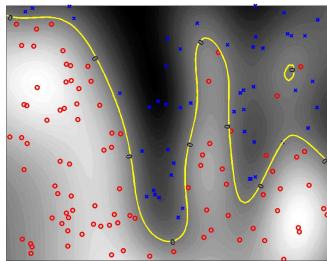
Kernel machine intuition



Kernel machine intuition



- (1) fit the data
- (2) regular solution
- (3) examples contribution
- (4) remove useless examples



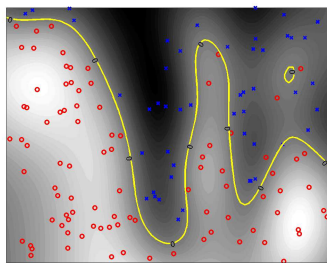
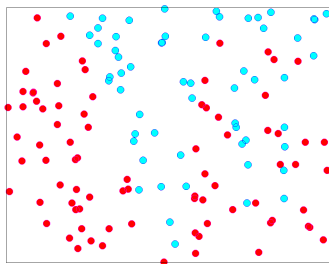
minimize R_{emp}

regularize

use **Kernels** \iff specify \mathcal{H}

be sparse

Kernel machine intuition



- (1) fit the data
- (2) regular solution
- (3) examples contribution
- (4) remove useless examples

minimize R_{emp}

regularize

use **Kernels** \iff specify \mathcal{H}

be sparse

Kernel machines: $\min_{f \in \mathcal{H}} \left(R_{\text{emp}} \text{ and } \mathcal{H} \text{ complexity} \right)$ with kernel based \mathcal{H}

universal, good statistically and computationally

Kernel machine

kernel as a dictionary

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i)$$

- α_i influence of example i
- $k(\mathbf{x}, \mathbf{x}_i)$ the kernel

depends on y_i
do NOT depend on y_i

Definition (Kernel)

Let \mathcal{X} be a non empty set (the input space).

A *kernel* is a function k from $\mathcal{X} \times \mathcal{X}$ onto \mathbb{R} .

$$k: \begin{array}{ccc} \mathcal{X} \times \mathcal{X} & \longmapsto & \mathbb{R} \\ \mathbf{s}, \mathbf{t} & \longrightarrow & k(\mathbf{s}, \mathbf{t}) \end{array}$$

Kernel machine

kernel as a dictionary

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i)$$

- α_i influence of example i
- $k(\mathbf{x}, \mathbf{x}_i)$ the kernel

depends on y_i
do NOT depend on y_i

Definition (Kernel)

Let \mathcal{X} be a non empty set (the input space).

A *kernel* is a function k from $\mathcal{X} \times \mathcal{X}$ onto \mathbb{R} .

$$\begin{array}{ccc} k : & \mathcal{X} \times \mathcal{X} & \mapsto \mathbb{R} \\ & \mathbf{s}, \mathbf{t} & \longrightarrow k(\mathbf{s}, \mathbf{t}) \end{array}$$

semi-parametric version: given the family $q_j(\mathbf{x})$, $j = 1, p$

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i) + \sum_{j=1}^p \beta_j q_j(\mathbf{x})$$

Machine à noyaux

Definition (Kernel machines)

$$\mathcal{A}((\mathbf{x}_i, y_i)_{i=1,n})(\mathbf{x}) = \psi\left(\sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i) + \sum_{j=1}^p \beta_j q_j(\mathbf{x})\right)$$

α et β : parameters to be estimated.

Exemples

$$\mathcal{A}(x) = \sum_{i=1}^n \alpha_i (x - x_i)_+^3 + \beta_0 + \beta_1 x \quad \text{splines}$$

$$\mathcal{A}(\mathbf{x}) = \text{sign}\left(\sum_{i \in I} \alpha_i \exp^{-\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{b}} + \beta_0\right) \quad \text{SVM}$$

$$\mathbb{P}(y|\mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{i \in I} \alpha_i \mathbb{1}_{\{y=y_i\}} (\mathbf{x}^\top \mathbf{x}_i + b)^2\right) \quad \text{exponential family}$$

Plan

1 Statistical learning and kernels

- The statistical learning problem
- **Kernels**
- Kernel and hypothesis set
- Functional differentiation in RKHS

In the beginning was the kernel...

Definition (Kernel)

a function of two variable k from $\mathcal{X} \times \mathcal{X}$ to \mathbb{R}

Definition (Positive kernel)

A kernel $k(s, t)$ on \mathcal{X} is said to be positive

- if it is symmetric: $k(s, t) = k(t, s)$
- and if for any finite positive integer n :

$$\forall \{\alpha_i\}_{i=1,n} \in \mathbb{R}, \forall \{\mathbf{x}_i\}_{i=1,n} \in \mathcal{X}, \quad \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

it is strictly positive if for $\alpha_i \neq 0$

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) > 0$$

Examples of positive kernels

the linear kernel: $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d$, $k(\mathbf{s}, \mathbf{t}) = \mathbf{s}^\top \mathbf{t}$

symetric: $\mathbf{s}^\top \mathbf{t} = \mathbf{t}^\top \mathbf{s}$

$$\begin{aligned} \text{positive: } \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \mathbf{x}_i^\top \mathbf{x}_j \\ &= \left(\sum_{i=1}^n \alpha_i \mathbf{x}_i \right)^\top \left(\sum_{j=1}^n \alpha_j \mathbf{x}_j \right) = \left\| \sum_{i=1}^n \alpha_i \mathbf{x}_i \right\|^2 \end{aligned}$$

the product kernel: $k(\mathbf{s}, \mathbf{t}) = g(\mathbf{s})g(\mathbf{t})$ for some $g : \mathbb{R}^d \rightarrow \mathbb{R}$,

symetric by construction

$$\begin{aligned} \text{positive: } \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j g(\mathbf{x}_i) g(\mathbf{x}_j) \\ &= \left(\sum_{i=1}^n \alpha_i g(\mathbf{x}_i) \right) \left(\sum_{j=1}^n \alpha_j g(\mathbf{x}_j) \right) = \left(\sum_{i=1}^n \alpha_i g(\mathbf{x}_i) \right)^2 \end{aligned}$$

k is positive \Leftrightarrow (its square root exists) $\Leftrightarrow k(\mathbf{s}, \mathbf{t}) = \langle \phi_{\mathbf{s}}, \phi_{\mathbf{t}} \rangle$

Example: finite kernel

let $\phi_j, j = 1, p$ be a finite dictionary of functions from \mathcal{X} to \mathbb{R} (polynomials, wavelets...)

the feature map and linear kernel

feature map:
$$\begin{aligned} \Phi : \mathcal{X} &\rightarrow \mathbb{R}^p \\ \mathbf{s} &\mapsto \Phi = (\phi_1(\mathbf{s}), \dots, \phi_p(\mathbf{s})) \end{aligned}$$

Linear kernel in the feature space:

$$k(\mathbf{s}, \mathbf{t}) = (\phi_1(\mathbf{s}), \dots, \phi_p(\mathbf{s}))^\top (\phi_1(\mathbf{t}), \dots, \phi_p(\mathbf{t}))$$

e.g. the quadratic kernel: $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d, \quad k(\mathbf{s}, \mathbf{t}) = (\mathbf{s}^\top \mathbf{t} + b)^2$

feature map:

$$\begin{aligned} \Phi : \mathbb{R}^d &\rightarrow \mathbb{R}^{p=1+d+\frac{d(d+1)}{2}} \\ \mathbf{s} &\mapsto \Phi = (1, \sqrt{2}s_1, \dots, \sqrt{2}s_j, \dots, \sqrt{2}s_d, s_1^2, \dots, s_j^2, \dots, s_d^2, \dots, \sqrt{2}s_i s_j, \dots) \end{aligned}$$

closed form kernel: the quadratic kernel

The quadratic dictionary in \mathbb{R}^d :

$$\begin{aligned}\Phi : \mathbb{R}^d &\rightarrow \mathbb{R}^{p=1+d+\frac{d(d+1)}{2}} \\ \mathbf{s} &\mapsto \Phi = (1, s_1, s_2, \dots, s_d, s_1^2, s_2^2, \dots, s_d^2, \dots, s_i s_j, \dots)\end{aligned}$$

in this case

$$\Phi(\mathbf{s})^\top \Phi(\mathbf{t}) = 1 + s_1 t_1 + s_2 t_2 + \dots + s_d t_d + s_1^2 t_1^2 + \dots + s_d^2 t_d^2 + \dots + s_i s_j t_i t_j + \dots$$

closed form kernel: the quadratic kernel

The quadratic dictionary in \mathbb{R}^d :

$$\begin{aligned}\Phi: \mathbb{R}^d &\rightarrow \mathbb{R}^{p=1+d+\frac{d(d+1)}{2}} \\ \mathbf{s} &\mapsto \Phi = (1, s_1, s_2, \dots, s_d, s_1^2, s_2^2, \dots, s_d^2, \dots, s_i s_j, \dots)\end{aligned}$$

in this case

$$\Phi(\mathbf{s})^\top \Phi(\mathbf{t}) = 1 + s_1 t_1 + s_2 t_2 + \dots + s_d t_d + s_1^2 t_1^2 + \dots + s_d^2 t_d^2 + \dots + s_i s_j t_i t_j + \dots$$

The quadratic kernel: $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d$, $k(\mathbf{s}, \mathbf{t}) = (\mathbf{s}^\top \mathbf{t} + 1)^2$
 $= 1 + 2\mathbf{s}^\top \mathbf{t} + (\mathbf{s}^\top \mathbf{t})^2$

computes the dot product of the reweighted dictionary:

$$\begin{aligned}\Phi: \mathbb{R}^d &\rightarrow \mathbb{R}^{p=1+d+\frac{d(d+1)}{2}} \\ \mathbf{s} &\mapsto \Phi = (1, \sqrt{2}s_1, \sqrt{2}s_2, \dots, \sqrt{2}s_d, s_1^2, s_2^2, \dots, s_d^2, \dots, \sqrt{2}s_i s_j, \dots)\end{aligned}$$

closed form kernel: the quadratic kernel

The quadratic dictionary in \mathbb{R}^d :

$$\begin{aligned}\Phi: \mathbb{R}^d &\rightarrow \mathbb{R}^{p=1+d+\frac{d(d+1)}{2}} \\ \mathbf{s} &\mapsto \Phi = (1, s_1, s_2, \dots, s_d, s_1^2, s_2^2, \dots, s_d^2, \dots, s_i s_j, \dots)\end{aligned}$$

in this case

$$\Phi(\mathbf{s})^\top \Phi(\mathbf{t}) = 1 + s_1 t_1 + s_2 t_2 + \dots + s_d t_d + s_1^2 t_1^2 + \dots + s_d^2 t_d^2 + \dots + s_i s_j t_i t_j + \dots$$

The quadratic kernel: $\mathbf{s}, \mathbf{t} \in \mathbb{R}^d$, $k(\mathbf{s}, \mathbf{t}) = (\mathbf{s}^\top \mathbf{t} + 1)^2$
 $= 1 + 2\mathbf{s}^\top \mathbf{t} + (\mathbf{s}^\top \mathbf{t})^2$

computes the dot product of the reweighted dictionary:

$$\begin{aligned}\Phi: \mathbb{R}^d &\rightarrow \mathbb{R}^{p=1+d+\frac{d(d+1)}{2}} \\ \mathbf{s} &\mapsto \Phi = (1, \sqrt{2}s_1, \sqrt{2}s_2, \dots, \sqrt{2}s_d, s_1^2, s_2^2, \dots, s_d^2, \dots, \sqrt{2}s_i s_j, \dots)\end{aligned}$$

$p = 1 + d + \frac{d(d+1)}{2}$ multiplications vs. $d + 1$
use kernel to save computation

Positive definite Kernel (PDK) algebra (closure)

if $k_1(\mathbf{s}, \mathbf{t})$ and $k_2(\mathbf{s}, \mathbf{t})$ are two positive kernels

- DPK are a convex cone:

$$\forall a_1 \in \mathbb{R}^+ \quad a_1 k_1(\mathbf{s}, \mathbf{t}) + k_2(\mathbf{s}, \mathbf{t})$$

- product kernel

$$k_1(\mathbf{s}, \mathbf{t}) k_2(\mathbf{s}, \mathbf{t})$$

proofs

- by linearity:

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j (a_1 k_1(i, j) + k_2(i, j)) = a_1 \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k_1(i, j) + \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k_2(i, j)$$

- assuming $\exists \psi_\ell$ s.t. $k_1(\mathbf{s}, \mathbf{t}) = \sum_{\ell} \psi_\ell(\mathbf{s}) \psi_\ell(\mathbf{t})$

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k_1(\mathbf{x}_i, \mathbf{x}_j) k_2(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \left(\sum_{\ell} \psi_\ell(\mathbf{x}_i) \psi_\ell(\mathbf{x}_j) k_2(\mathbf{x}_i, \mathbf{x}_j) \right) \\ &= \sum_{\ell} \sum_{i=1}^n \sum_{j=1}^n (\alpha_i \psi_\ell(\mathbf{x}_i)) (\alpha_j \psi_\ell(\mathbf{x}_j)) k_2(\mathbf{x}_i, \mathbf{x}_j) \end{aligned}$$

Kernel engineering: building PDK

- for any polynomial with positive coef. ϕ from \mathbb{R} to \mathbb{R}

$$\phi(k(s, t))$$

- if Ψ is a function from \mathbb{R}^d to \mathbb{R}^d

$$k(\Psi(s), \Psi(t))$$

- if φ from \mathbb{R}^d to \mathbb{R}^+ , is minimum in 0

$$k(s, t) = \varphi(s + t) - \varphi(s - t)$$

- convolution of two positive kernels is a positive kernel

$$K_1 \star K_2$$

Example : the Gaussian kernel is a PDK

$$\begin{aligned}\exp(-\|s - t\|^2) &= \exp(-\|s\|^2 - \|t\|^2 - 2s^\top t) \\ &= \exp(-\|s\|^2) \exp(-\|t\|^2) \exp(2s^\top t)\end{aligned}$$

- $s^\top t$ is a PDK and function \exp as the limit of positive series expansion, so $\exp(2s^\top t)$ is a PDK
- $\exp(-\|s\|^2) \exp(-\|t\|^2)$ is a PDK as a product kernel
- the product of two PDK is a PDK

an attempt at classifying PD kernels

- stationary kernels, (also called translation invariant):

$$k(s, t) = k_s(s - t)$$

- radial (isotropic) gaussian: $\exp\left(-\frac{r^2}{b}\right)$, $r = \|s - t\|$
- with compact support

c.s. Matérn : $\max\left(0, 1 - \left(\frac{r}{b}\right)^\kappa\right) \frac{r}{b}^\kappa B_\kappa\left(\frac{r}{b}\right)$, $\kappa \geq (d + 1)/2$

- locally stationary kernels: $k(s, t) = k_1(s + t)k_s(s - t)$
 K_1 is a non negative function and K_2 a radial kernel.

- non stationary (projective kernels):

$$k(s, t) = k_p(s^\top t)$$

- separable kernels $k(s, t) = k_1(s)k_2(t)$ with k_1 and $k_2(t)$ PDK
in this case $K = k_1 k_2^\top$ where $k_1 = (k_1(\mathbf{x}_1), \dots, k_1(\mathbf{x}_n))$

some examples of PD kernels...

type	name	$k(s, t)$
radial	gaussian	$\exp\left(-\frac{r^2}{b}\right), \quad r = \ s - t\ $
radial	laplacian	$\exp(-r/b)$
radial	rational	$1 - \frac{r^2}{r^2 + b}$
radial	loc. gauss.	$\max\left(0, 1 - \frac{r}{3b}\right)^d \exp(-\frac{r^2}{b})$
non stat.	χ^2	$\exp(-r/b), \quad r = \sum_k \frac{(s_k - t_k)^2}{s_k + t_k}$
projective	polynomial	$(s^\top t)^p$
projective	affine	$(s^\top t + b)^p$
projective	cosine	$s^\top t / \ s\ \ t\ $
projective	correlation	$\exp\left(\frac{s^\top t}{\ s\ \ t\ } - b\right)$

Most of the kernels depends on a quantity b called the bandwidth

the importance of the Kernel bandwidth

for the affine Kernel: Bandwidth = bias

$$k(\mathbf{s}, \mathbf{t}) = (\mathbf{s}^\top \mathbf{t} + b)^p = b^p \left(\frac{\mathbf{s}^\top \mathbf{t}}{b} + 1 \right)^p$$

for the gaussian Kernel: Bandwidth = influence zone

$$k(\mathbf{s}, \mathbf{t}) = \frac{1}{Z} \exp \left(-\frac{\|\mathbf{s} - \mathbf{t}\|^2}{2\sigma^2} \right) \quad b = 2\sigma^2$$

the importance of the Kernel bandwidth

for the affine Kernel: Bandwidth = biais

$$k(\mathbf{s}, \mathbf{t}) = (\mathbf{s}^\top \mathbf{t} + b)^p = b^p \left(\frac{\mathbf{s}^\top \mathbf{t}}{b} + 1 \right)^p$$

for the gaussian Kernel: Bandwidth = influence zone

$$k(\mathbf{s}, \mathbf{t}) = \frac{1}{Z} \exp \left(-\frac{\|\mathbf{s} - \mathbf{t}\|^2}{2\sigma^2} \right) \quad b = 2\sigma^2$$

Illustration

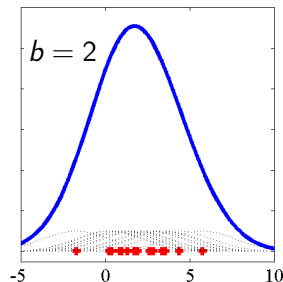
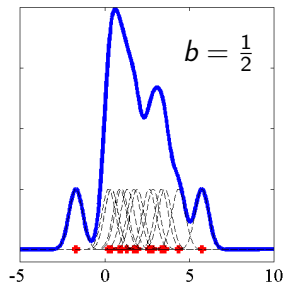
1 d density estimation

+ data

(x_1, x_2, \dots, x_n)

– Parzen estimate

$$\hat{\mathbb{P}}(x) = \frac{1}{Z} \sum_{i=1}^n k(x, x_i)$$



kernels for objects and structures

kernels on histograms and probability distributions

kernel on strings

- spectral string kernel
- using sub sequences
- similarities by alignements

$$k(\mathbf{s}, \mathbf{t}) = \sum_u \phi_u(\mathbf{s}) \phi_u(\mathbf{t})$$

$$k(\mathbf{s}, \mathbf{t}) = \sum_{\pi} \exp(\beta(\mathbf{s}, \mathbf{t}, \pi))$$

kernels on graphs

- the pseudo inverse of the (regularized) graph Laplacian

$$L = D - A \quad A \text{ is the adjacency matrix } D \text{ the degree matrix}$$

- diffusion kernels
- subgraph kernel convolution (using random walks)

$$\frac{1}{Z(b)} \exp^{bL}$$

and kernels on HMM, automata, dynamical system...

Multiple kernel

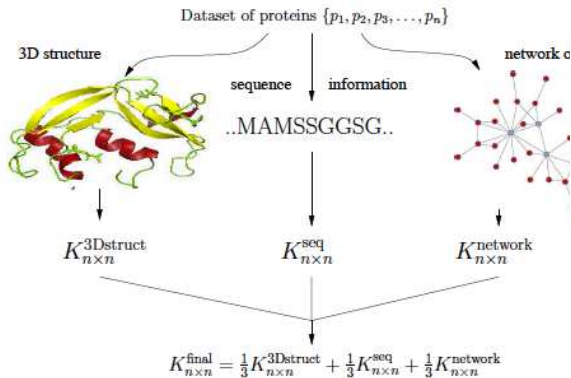


Figure 2: A dataset of proteins can be regarded in (at least) three different ways: as 3D structures, a dataset of sequences and a set of nodes in a network which in turn can be represented as a graph. A different kernel matrix can be extracted from each datatype, using known shapes, strings and graphs. The resulting kernels can then be combined together using different weights, as is the case above where a simple average is considered, or estimated, as the subject of Section 5.2.

Gram matrix

Definition (Gram matrix)

let $k(\mathbf{s}, \mathbf{t})$ be a positive kernel on \mathcal{X} and $(\mathbf{x}_i)_{i=1,n}$ a sequence on \mathcal{X} . the Gram matrix is the square K of dimension n and of general term $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

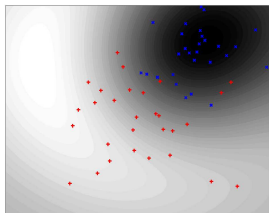
practical trick to check kernel positivity:

K is positive $\Leftrightarrow \lambda_i > 0$ its eigenvalues are positives: if $K\mathbf{u}_i = \lambda_i\mathbf{u}_i; i = 1, n$

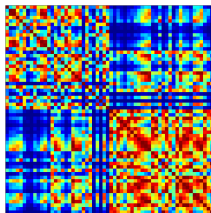
$$\mathbf{u}_i^\top K \mathbf{u}_i = \lambda_i \mathbf{u}_i^\top \mathbf{u}_i = \lambda_i$$

matrix K is the one to be used

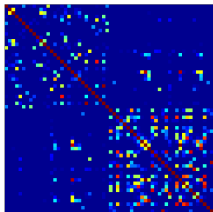
Examples of Gram matrices with different bandwidth



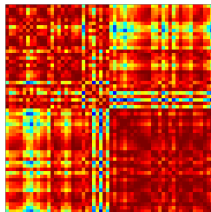
raw data



Gram matrix for $b = 2$



$b = .5$



$b = 10$

different point of view about kernels

kernel and scalar product

$$k(\mathbf{s}, \mathbf{t}) = \langle \phi(\mathbf{s}), \phi(\mathbf{t}) \rangle_{\mathcal{H}}$$

kernel and distance

$$d(\mathbf{s}, \mathbf{t})^2 = k(\mathbf{s}, \mathbf{s}) + k(\mathbf{t}, \mathbf{t}) - 2k(\mathbf{s}, \mathbf{t})$$

kernel and covariance: a positive matrix is a covariance matrix

$$\mathbb{P}(\mathbf{f}) = \frac{1}{Z} \exp\left(-\frac{1}{2}(\mathbf{f} - \mathbf{f}_0)^\top K^{-1}(\mathbf{f} - \mathbf{f}_0)\right)$$

$$\text{if } \mathbf{f}_0 = \mathbf{0} \text{ and } \mathbf{f} = K\alpha, \mathbb{P}(\alpha) = \frac{1}{Z} \exp\left(-\frac{1}{2}\alpha^\top K\alpha\right)$$

Kernel and regularity (green's function)

$$k(\mathbf{s}, \mathbf{t}) = P^* P \delta_{\mathbf{s}-\mathbf{t}} \quad \text{for some operator } P \quad (\text{e.g. some differential})$$

Let's summarize

- positive kernels
- there is a lot of them
- can be rather complex
- 2 classes: radial / projective
- the bandwidth matters (more than the kernel itself)
- the Gram matrix summarize the pairwise comparisons

Roadmap

1 Statistical learning and kernels

- The statistical learning problem
- Kernels
- Kernel and hypothesis set
- Functional differentiation in RKHS

From kernel to functions

$$\mathcal{H}_0 = \left\{ f \mid m_f < \infty; f_j \in \mathbb{R}; t_j \in \mathcal{X}, f(\mathbf{x}) = \sum_{j=1}^{m_f} f_j k(\mathbf{x}, t_j) \right\}$$

let define the bilinear form ($g(\mathbf{x}) = \sum_{i=1}^{m_g} g_i k(\mathbf{x}, s_i)$) :

$$\forall f, g \in \mathcal{H}_0, \langle f, g \rangle_{\mathcal{H}_0} = \sum_{j=1}^{m_f} \sum_{i=1}^{m_g} f_j g_i k(t_j, s_i)$$

Evaluation functional: $\forall \mathbf{x} \in \mathcal{X}$

$$f(\mathbf{x}) = \langle f(\bullet), k(\mathbf{x}, \bullet) \rangle_{\mathcal{H}_0}$$

from k to \mathcal{H}

for any positive kernel, a hypothesis set can be constructed $\mathcal{H} = \overline{\mathcal{H}_0}$ with its metric

RKHS

Definition (reproducing kernel Hilbert space (RKHS))

a Hilbert space \mathcal{H} embedded with the inner product $\langle \bullet, \bullet \rangle_{\mathcal{H}}$ is said to be with reproducing kernel if it exists a positive kernel k such that

$$\begin{aligned}\forall s \in \mathcal{X}, \quad k(\bullet, s) &\in \mathcal{H} \\ \forall f \in \mathcal{H}, \quad f(s) &= \langle f(\bullet), k(s, \bullet) \rangle_{\mathcal{H}}\end{aligned}$$

Beware: $f = f(\bullet)$ is a function while $f(s)$ is the real value of f at point s

positive kernel \Leftrightarrow RKHS

- any function in \mathcal{H} is pointwise defined
- defines the inner product
- it defines the **regularity** (smoothness) of the hypothesis set

Exercise: let $f(\bullet) = \sum_{i=1}^n \alpha_i k(\bullet, x_i)$. Show that $\|f\|_{\mathcal{H}}^2 = \alpha^\top K \alpha$

Remark about the regularity matter: from \mathcal{H} to k

How to build \mathcal{H} together with k ?

let Γ_t be a family in L^2 , $t \in \mathcal{X}$ (e.g. $\Gamma_t(s) = \mathbb{I}_{\{s \leq t\}}$; $\mathcal{X} = [0, 1]$)

define the following mapping

$$S_t: L^2 \rightarrow \mathbb{R}$$

$$\begin{aligned} g &\mapsto S_t(g) = \int g(x) \Gamma_t(x) dx \\ &= \int_0^1 g(x) \mathbb{I}_{\{x \leq t\}} dx \end{aligned}$$

$$f(t) = S_t(g) = \langle g, \Gamma_t \rangle_{L^2}$$

$$f(t) = \int_0^t g(x) dx = G(t)$$

the following RKHS can be constructed

$$\mathcal{H} = \text{Im}(S) \quad \langle f_1, f_2 \rangle_{\mathcal{H}} = \langle g_1, g_2 \rangle_{L^2} \quad k(s, t) = \langle \Gamma_s, \Gamma_t \rangle_{L^2}$$

the reproducing property is verified:

$$\langle f(\cdot), k(\cdot, t) \rangle_{\mathcal{H}} = \langle g, \Gamma_t \rangle_{L^2} = f(t)$$

Γ_x examples and associated kernels

	$\Gamma_x(u)$	$K(x, y)$
Cameron Martin	$\mathbb{I}_{\{x \leq \frac{u}{d}\}}$	$\min(x, y)$
Polynomial	$e_0(u) + \sum_{i=1} x_i e_i(u)$	$\mathbf{x}^\top \mathbf{y} + 1$
Gaussian	$\frac{1}{Z} \exp^{-\frac{(x-u)^2}{2}}$	$\frac{1}{Z'} \exp^{-\frac{(x-y)^2}{4}}$

$\{e_i\}_{i=1,d}$ are a finite sub sample of an orthonormal basis in L^2

Γ_x examples and associated kernels

	$\Gamma_x(u)$	$K(x, y)$
Cameron Martin	$\mathbb{I}_{\{x \leq u\}}$	$\min(x, y)$
Polynomial	$e_0(u) + \sum_{i=1}^d x_i e_i(u)$	$\mathbf{x}^\top \mathbf{y} + 1$
Gaussian	$\frac{1}{Z} \exp^{-\frac{(x-u)^2}{2}}$	$\frac{1}{Z'} \exp^{-\frac{(x-y)^2}{4}}$

$\{e_i\}_{i=1,d}$ are a finite sub sample of an orthonormal basis in L^2

the Cameron Martin operator: $\forall f \in \mathcal{H}, \exists g \in L^2$ such that

$$f(t) = (Sg)(t) = \int \Gamma_t(u) g(u) du = \int \mathbb{I}_{\{t \leq u\}} g(u) du = G(t)$$

$$\|f\|_{\mathcal{H}} = \|Sg\|_{\mathcal{H}} = \|g\|_{L^2}^2 = \|f'\|_{L^2}^2 \quad \text{generalized differential}$$

$$\text{integral} \quad S = P^{-1} \quad \text{differential}$$

other kernels (what really matters)

- finite kernels

$$k(\mathbf{s}, \mathbf{t}) = (\phi_1(\mathbf{s}), \dots, \phi_p(\mathbf{s}))^\top (\phi_1(\mathbf{t}), \dots, \phi_p(\mathbf{t}))$$

- Mercer kernels

positive on a compact set

\Leftrightarrow

$$k(\mathbf{s}, \mathbf{t}) = \sum_{j=1}^p \lambda_j \phi_j(\mathbf{s}) \phi_j(\mathbf{t})$$

- positive kernels

- positive semi-definite

- conditionnaly positive (for some functions p_j)

$$\forall \{\mathbf{x}_i\}_{i=1,n}, \forall \alpha_i, \sum_i^n \alpha_i p_j(\mathbf{x}_i) = 0; \quad j = 1, p, \quad \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

- symetric non positive

$$k(\mathbf{s}, \mathbf{t}) = \tanh(\mathbf{s}^\top \mathbf{t} + \alpha_0)$$

- non symetric – non positive

the key property: $\nabla_{J_t}(f) = k(t, \cdot)$ holds

The kernel map

- observation: $\mathbf{x} = (x_1, \dots, x_j, \dots, x_d)^\top$
 - ▶ $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} = \langle \mathbf{w}, \mathbf{x} \rangle_{\mathbb{R}^d}$
- feature map: $\mathbf{x} \longrightarrow \Phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \dots, \phi_j(\mathbf{x}), \dots, \phi_p(\mathbf{x}))^\top$
 - ▶ $\Phi : \mathbb{R}^d \mapsto \mathbb{R}^p$
 - ▶ $f(\mathbf{x}) = \mathbf{w}^\top \Phi(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle_{\mathbb{R}^p}$
- kernel dictionary: $\mathbf{x} \longrightarrow \mathbf{k}(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}_1), \dots, k(\mathbf{x}, \mathbf{x}_i), \dots, k(\mathbf{x}, \mathbf{x}_n))^\top$
 - ▶ $\mathbf{k} : \mathbb{R}^d \mapsto \mathbb{R}^n$
 - ▶ $f(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i) = \langle \alpha, \mathbf{k}(\mathbf{x}) \rangle_{\mathbb{R}^n}$
- **kernel map:** $\mathbf{x} \longrightarrow k(\bullet, \mathbf{x}) \quad p = \infty$
 - ▶ $f(\mathbf{x}) = \langle f(\bullet), K(\bullet, \mathbf{x}) \rangle_{\mathcal{H}}$

Roadmap

- 1 Statistical learning and kernels
 - The statistical learning problem
 - Kernels
 - Kernel and hypothesis set
 - Functional differentiation in RKHS

Functional differentiation in RKHS

Let J be a functional

$$\begin{array}{ll} J: \mathcal{H} \rightarrow \mathbb{R} \\ f \mapsto J(f) \end{array} \quad \text{examples:} \quad J_1(f) = \|f\|_{\mathcal{H}}^2, J_2(f) = f(\mathbf{x}),$$

J directional derivative in direction g at point f

$$dJ(f, g) = \lim_{\varepsilon \rightarrow 0} \frac{J(f + \varepsilon g) - J(f)}{\varepsilon}$$

Gradient $\nabla_J(f)$

$$\begin{array}{ll} \nabla_J: \mathcal{H} \rightarrow \mathcal{H} \\ f \mapsto \nabla_J(f) \end{array} \quad \text{si} \quad dJ(f, g) = \langle \nabla_J(f), g \rangle_{\mathcal{H}}$$

exercise: find out $\nabla_{J_1}(f)$ et $\nabla_{J_2}(f)$

Hint

$$dJ(f, g) = \left. \frac{dJ(f + \varepsilon g)}{d\varepsilon} \right|_{\varepsilon=0}$$

Solution

$$\begin{aligned} dJ_1(f, g) &= \lim_{\varepsilon \rightarrow 0} \frac{\|f + \varepsilon g\|^2 - \|f\|^2}{\varepsilon} \\ &= \lim_{\varepsilon \rightarrow 0} \frac{\|f\|^2 + \varepsilon^2 \|g\|^2 + 2\varepsilon \langle f, g \rangle_{\mathcal{H}} - \|f\|^2}{\varepsilon} \\ &= \lim_{\varepsilon \rightarrow 0} \varepsilon \|g\|^2 + 2 \langle f, g \rangle_{\mathcal{H}} \quad \Leftrightarrow \quad \nabla_{J_1}(f) = 2f \\ &= \langle 2f, g \rangle_{\mathcal{H}} \end{aligned}$$

$$\begin{aligned} dJ_2(f, g) &= \lim_{\varepsilon \rightarrow 0} \frac{f(\mathbf{x}) + \varepsilon g(\mathbf{x}) - f(\mathbf{x})}{\varepsilon} \\ &= g(\mathbf{x}) \quad \Leftrightarrow \quad \nabla_{J_2}(f) = k(\mathbf{x}, \cdot) \\ &= \langle k(\mathbf{x}, \cdot), g \rangle_{\mathcal{H}} \end{aligned}$$

Solution

$$\begin{aligned} dJ_1(f, g) &= \lim_{\varepsilon \rightarrow 0} \frac{\|f + \varepsilon g\|^2 - \|f\|^2}{\varepsilon} \\ &= \lim_{\varepsilon \rightarrow 0} \frac{\|f\|^2 + \varepsilon^2 \|g\|^2 + 2\varepsilon \langle f, g \rangle_{\mathcal{H}} - \|f\|^2}{\varepsilon} \\ &= \lim_{\varepsilon \rightarrow 0} \varepsilon \|g\|^2 + 2\langle f, g \rangle_{\mathcal{H}} \quad \Leftrightarrow \quad \nabla_{J_1}(f) = 2f \\ &= \langle 2f, g \rangle_{\mathcal{H}} \end{aligned}$$

$$\begin{aligned} dJ_2(f, g) &= \lim_{\varepsilon \rightarrow 0} \frac{f(\mathbf{x}) + \varepsilon g(\mathbf{x}) - f(\mathbf{x})}{\varepsilon} \\ &= g(\mathbf{x}) \quad \Leftrightarrow \quad \nabla_{J_2}(f) = k(\mathbf{x}, \cdot) \\ &= \langle k(\mathbf{x}, \cdot), g \rangle_{\mathcal{H}} \end{aligned}$$

$$\text{Minimize}_{f \in \mathcal{H}} J(f) \quad \Leftrightarrow \quad \forall g \in \mathcal{H}, dJ(f, g) = 0 \quad \Leftrightarrow \quad \nabla_J(f) = 0$$

Subdifferential in a RKHS \mathcal{H}

Definition (Sub gradient)

a subgradient of $J : \mathcal{H} \mapsto \mathbb{R}$ at f_0 is any function $g \in \mathcal{H}$ such that

$$\forall f \in \mathcal{V}(f_0), \quad J(f) \geq J(f_0) + \langle g, (f - f_0) \rangle_{\mathcal{H}}$$

Definition (Subdifferential)

$\partial J(f)$, the subdifferential of J at f is the set of all subgradients of J at f .

$$\mathcal{H} = \mathbb{R} \quad J_3(x) = |x| \quad \partial J_3(0) = \{g \in \mathbb{R} \mid -1 \leq g \leq 1\}$$

$$\mathcal{H} = \mathbb{R} \quad J_4(x) = \max(0, 1 - x) \quad \partial J_4(1) = \{g \in \mathbb{R} \mid -1 \leq g \leq 0\}$$

Theorem (Chain rule for linear Subdifferential)

Let T be a linear operator $\mathcal{H} \mapsto \mathbb{R}$ and φ a function from \mathbb{R} to \mathbb{R}

If $J(f) = \varphi(Tf)$

Then $\partial J(f) = \{T^*g \mid g \in \partial\varphi(Tf)\}$, where T^* denotes T 's adjoint operator

example of subdifferential in \mathcal{H}

evaluation operator and its adjoint

$$T : \mathcal{H} \longrightarrow \mathbb{R}^n$$

$$f \longmapsto Tf = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^\top$$

$$T^* : \mathbb{R}^n \longrightarrow \mathcal{H}$$

$$\alpha \longmapsto T^* \alpha$$

build the adjoint $\langle Tf, \alpha \rangle_{\mathbb{R}^n} = \langle f, T^* \alpha \rangle_{\mathcal{H}}$

example of subdifferential in \mathcal{H}
 evaluation operator and its adjoint

$$\begin{aligned} T : \mathcal{H} &\longrightarrow \mathbb{R}^n \\ f &\longmapsto Tf = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^\top \end{aligned}$$

$$\begin{aligned} T^* : \mathbb{R}^n &\longrightarrow \mathcal{H} \\ \alpha &\longmapsto T^*\alpha = \sum_{i=1}^n \alpha_i k(\bullet, \mathbf{x}_i) \end{aligned}$$

build the adjoint $\langle Tf, \alpha \rangle_{\mathbb{R}^n} = \langle f, T^*\alpha \rangle_{\mathcal{H}}$

$$\begin{aligned} \langle Tf, \alpha \rangle_{\mathbb{R}^n} &= \sum_{i=1}^n f(\mathbf{x}_i) \alpha_i \\ &= \sum_{i=1}^n \langle f(\bullet), k(\bullet, \mathbf{x}_i) \rangle_{\mathcal{H}} \alpha_i \\ &= \langle f(\bullet), \underbrace{\sum_{i=1}^n \alpha_i k(\bullet, \mathbf{x}_i)}_{T^*\alpha} \rangle_{\mathcal{H}} \end{aligned}$$

example of subdifferential in \mathcal{H}
 evaluation operator and its adjoint

$$\begin{aligned} T : \mathcal{H} &\longrightarrow \mathbb{R}^n \\ f &\longmapsto Tf = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^\top \end{aligned}$$

$$\begin{aligned} T^* : \mathbb{R}^n &\longrightarrow \mathcal{H} \\ \alpha &\longmapsto T^*\alpha = \sum_{i=1}^n \alpha_i k(\bullet, \mathbf{x}_i) \end{aligned}$$

build the adjoint $\langle Tf, \alpha \rangle_{\mathbb{R}^n} = \langle f, T^*\alpha \rangle_{\mathcal{H}}$

$$\begin{aligned} \langle Tf, \alpha \rangle_{\mathbb{R}^n} &= \sum_{i=1}^n f(\mathbf{x}_i) \alpha_i \\ &= \sum_{i=1}^n \langle f(\bullet), k(\bullet, \mathbf{x}_i) \rangle_{\mathcal{H}} \alpha_i \\ &= \langle f(\bullet), \underbrace{\sum_{i=1}^n \alpha_i k(\bullet, \mathbf{x}_i)}_{T^*\alpha} \rangle_{\mathcal{H}} \end{aligned}$$

$$\begin{aligned} TT^* : \mathbb{R}^n &\longrightarrow \mathbb{R}^n \\ \alpha &\longmapsto TT^*\alpha = \sum_{j=1}^n \alpha_j k(\mathbf{x}_j, \mathbf{x}_i) \\ &= K\alpha \end{aligned}$$

example of subdifferential in \mathcal{H}

evaluation operator and its adjoint

$$\begin{aligned} T : \mathcal{H} &\longrightarrow \mathbb{R}^n \\ f &\longmapsto Tf = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^{\top} \end{aligned}$$

$$\begin{aligned} T^* : \mathbb{R}^n &\longrightarrow \mathcal{H} \\ \alpha &\longmapsto T^*\alpha = \sum_{i=1}^n \alpha_i k(\bullet, \mathbf{x}_i) \end{aligned}$$

build the adjoint $\langle Tf, \alpha \rangle_{\mathbb{R}^n} = \langle f, T^*\alpha \rangle_{\mathcal{H}}$

$$\begin{aligned} \langle Tf, \alpha \rangle_{\mathbb{R}^n} &= \sum_{i=1}^n f(\mathbf{x}_i) \alpha_i \\ &= \sum_{i=1}^n \langle f(\bullet), k(\bullet, \mathbf{x}_i) \rangle_{\mathcal{H}} \alpha_i \\ &= \langle f(\bullet), \underbrace{\sum_{i=1}^n \alpha_i k(\bullet, \mathbf{x}_i)}_{T^*\alpha} \rangle_{\mathcal{H}} \end{aligned}$$

$$\begin{aligned} TT^* : \mathbb{R}^n &\longrightarrow \mathbb{R}^n \\ \alpha &\longmapsto TT^*\alpha = \sum_{j=1}^n \alpha_j k(\mathbf{x}_j, \mathbf{x}_i) \\ &= K\alpha \end{aligned}$$

Example of subdifferentials

x given $J_5(f) = |f(x)|$

$\partial J_5(f_0) = \{g(\bullet) = \alpha k(\bullet, \mathbf{x}) ; -1 < \alpha < 1\}$

x given $J_6(f) = \max(0, 1 - f(x))$

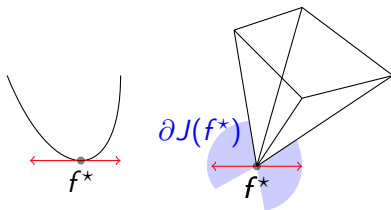
$\partial J_6(f_1) = \{g(\bullet) = \alpha k(\bullet, \mathbf{x}) ; -1 < \alpha < 0\}$

Optimal conditions

Theorem (Fermat optimality criterion)

When $J(f)$ is convex, f^\star is a stationary point of problem $\min_{f \in \mathcal{H}} J(f)$

If and only if $0 \in \partial J(f^\star)$



exercice: find for a given $y \in \mathbb{R}$ (from Obozinski)

$$\min_{x \in \mathbb{R}} \frac{1}{2}(x - y)^2 + \lambda|x|$$

Let's summarize

- positive kernels \Leftrightarrow RKHS $= \mathcal{H}$ \Leftrightarrow regularity $\|f\|_{\mathcal{H}}^2$
- the key property: $\nabla_{J_t}(f) = k(t, \cdot)$ holds not only for positive kernels
 $f(\mathbf{x}_i)$ exists (pointwise defined functions)
- universal consistency in RKHS
- the Gram matrix summarize the pairwise comparisons